AD-A096 716 PRINCETON UNIV NJ DEPT OF STATISTICS F/G 12/1
EYE-FITTING OF STRAIGHT LINES.(U)
JAN 81 F MOSTELLER, A F SIEGEL, E TRAPIDO DAAG29-79-C-0205
UNCLASSIFIED TR-183-SER-2 ARO-16669.5-M NL

END
DATE FILMED
4-81
DTIC

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LEVEL II

# REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

| | |
|---|---|
| 1. REPORT NUMBER: 16669.5-M | 2. GOVT ACCESSION NO.: AD-A096716 |
| 3. RECIPIENT'S CATALOG NUMBER: | |
| 4. TITLE (and Subtitle): Eye-Fitting of Straight Lines | 5. TYPE OF REPORT & PERIOD COVERED: Technical rept. |
| | 6. PERFORMING ORG. REPORT NUMBER: |
| 7. AUTHOR(s): Frederick Mosteller, Andrew F. Siegel, Edward Trapido, Cleo Youtz | 8. CONTRACT OR GRANT NUMBER(s): DAAG29-79-C-0205 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS: Princeton University, Princeton, NJ 08544 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: TR-123-SER 2 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709 | 12. REPORT DATE: 11 Jan 81 |
| | 13. NUMBER OF PAGES: 9 |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) | 15. SECURITY CLASS. (of this report): Unclassified |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DTIC ELECTE MAR 24 1981 E

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

least squares
regression
subjective fitting
principal components

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Because little is known about properties of lines fitted by eye, we designed and carried out an empirical investigation. Inexperienced graduate and post-doctoral students instructed to locate a line for estimating y from x for four sets of points tended to choose slopes near that of the first principal component (major axis) of the data and their lines passed close to the centroids. Students had a slight tendency to choose consistently either steeper or shallower slopes for all sets of data.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

# EYE-FITTING OF STRAIGHT LINES

by

Frederick Mosteller
Andrew F. Siegel
Edward Trapido
and
Cleo Youtz*

Technical Report No. 183, Series 2
Department of Statistics
Princeton University

January 1981

*This work was facilitated by National Science Foundation Grant No. NSF SOC75-15702 and Public Health Service Grant No. 5-D04-AH01698-02, both to Harvard University, and U.S. Army Research Grant DAAG29-79-C-0205 to Princeton University.

# EYE-FITTING OF STRAIGHT LINES

by

Frederick Mosteller, Andrew F. Siegel, Edward Trapido, and Cleo Youtz*

## A B S T R A C T

Because little is known about properties of lines fitted by eye, we designed and carried out an empirical investigation. Inexperienced graduate and post-doctoral students instructed to locate a line for estimating $y$ from $x$ for four sets of points tended to choose slopes near that of the first principal component (major axis) of the data and their lines passed close to the centroids. Students had a slight tendency to choose consistently either steeper or shallower slopes for all sets of data.

KEY WORDS: Least squares; Regression; Subjective fitting; Principal components.

---

*Frederick Mosteller is Roger I. Lee Professor, Department of Biostatistics, School of Public Health, Harvard University, 677 Huntington Avenue, Boston, MA, 02115. Andrew F. Siegel is Assistant Professor, Department of Statistics, Princeton University, Princeton, NJ, 08544. Edward Trapido is Teaching Fellow, Department of Epidemiology, School of Public Health, Harvard University, Boston, MA, 02115. Cleo Youtz is mathematical assistant, Department of Statistics, Harvard University, Cambridge, MA, 02138.

## 1. INTRODUCTION

The properties of least squares and other computed lines are well understood, but surprisingly little is known about the commonly used method of fitting by eye. This method involves maneuvering a string, black thread, or ruler until the fit seems satisfactory, and then drawing a line. We report one systematic investigation of eye-fitting of lines.

Students fitted lines by eye to four sets of points given in an experimental design to help us discover the properties of their fitted lines and whether order of fitting or practice made a difference. Other populations of subjects may produce different results. These sets of data were not unusual in curvature or in having outlying points or patterns. Thus additional populations of data sets could profitably be investigated.

The principal quantitative reference on fitting straight lines by eye is Finney (1951). He found that a mathematical iteration starting with slopes provided by scientists, inexperienced with probit analysis, gave satisfactory approximations to the relative potency in a bioassay.

## 2. METHOD

We conducted this experiment in a class of graduate and post-doctoral students in introductory biostatistics. Most students had not studied statistics before and had not yet been shown formal methods for fitting lines. The idea of using a regression line fitted to a set of points to estimate the vertical value, $y$, from the horizontal value, $x$, had been illustrated in a previous class session.

Each student was given the same set of four scatter diagrams and an 8½x11 inch transparency with a straight line etched completely across the middle. Students moved the transparency over the scatter diagram until satisfied with the fit of the etched line, and then marked an x on the scatter diagram at each end of the line. This transparency method is preferable to the black-thread method, which requires three hands.

The four scatter diagrams were labeled S for standard, F for fat, V for vertical, and N for negative; these are shown in Figure A. Data sets S, F, and V are linear transformations of each other, so that F has more vertical error than S and V has a steeper slope than S. Data sets S, F, and V come from a table of random numbers in Beyer (1971), whereas data set N is a linear transformation of the fiber strength data on page 224 of Dunn and Clark (1974).

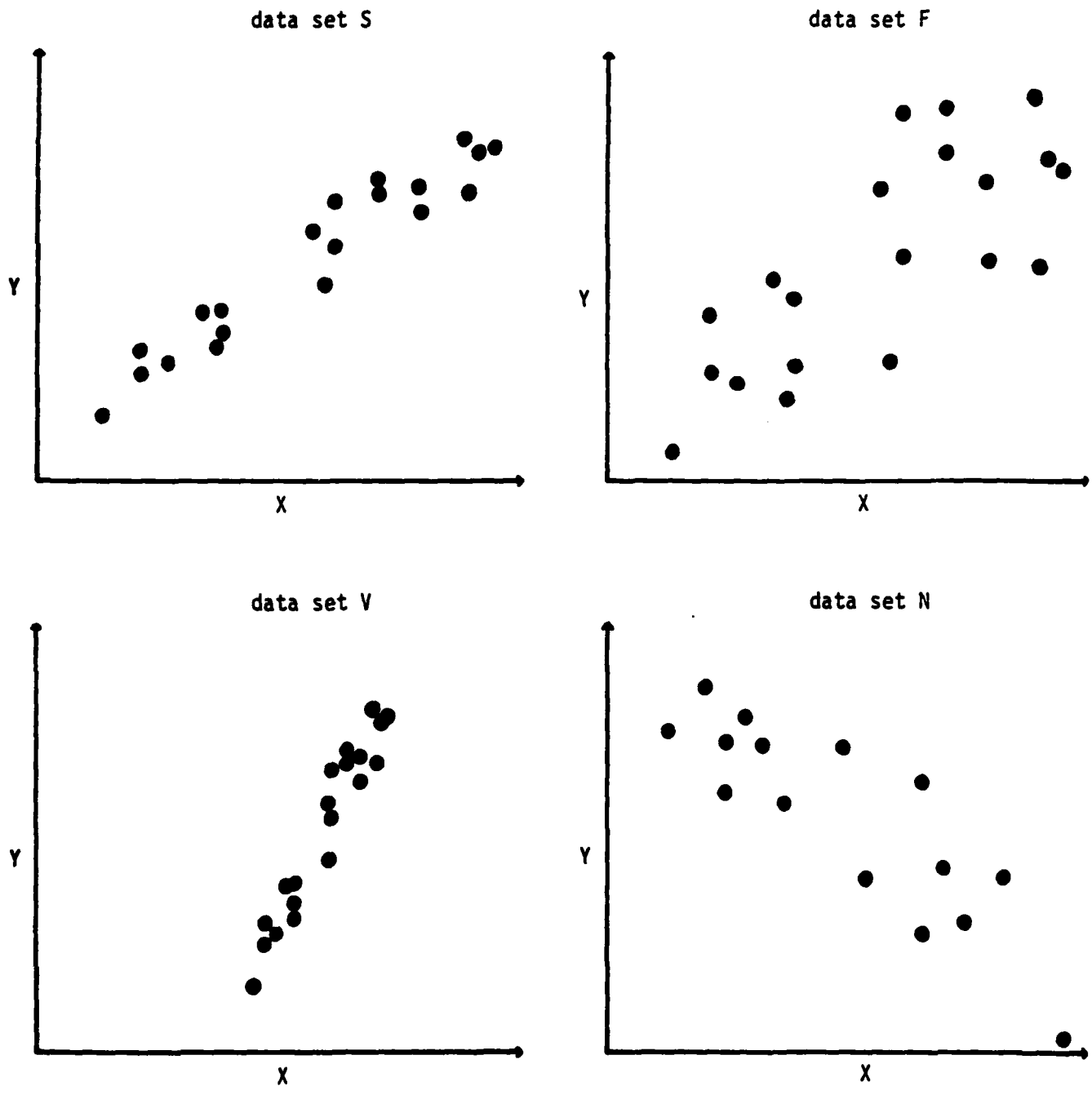


FIGURE A.

The data sets S, F, V, and N

To assess the effect of the order of presentation, we used a Latin square design with packets stapled in four different orders: SNFV, NSVF, FVSN, and VFNS. We distributed them systematically in that sequence so that students sitting side by side had different kinds of packets. We laid out on desks before class 175 packets and collected 153 at the end of the hour.

## 3. RESULTS

Table 1 summarizes the averages, variabilities, and actual (least squares) values for the slope and intercept of each data set. We have reported medians and interquartile ranges to reduce the effect of the few outlying values. The y-intercept at $\bar{x}$ measures the height of a line as well as does the y-intercept at zero, and is less correlated with the slope. To get Table 1, we pooled results from the four orders of presentation because we found no trend in the differences due to order.

TABLE 1.
Averages, Variabilities, and Actual Values for Slopes and Intercepts

| | S | F | V | N |
|---|---|---|---|---|
| Slope | | | | |
| median (interquartile range) | .70(.04) | .84(.14) | 2.07(.14) | -.73(.20) |
| actual least squares | | | | |
| regression | .66 | .66 | 1.98 | -.70 |
| principal component | .68 | .82 | 2.11 | -.79 |
| y-intercept at $\bar{x}$ | | | | |
| median (interquartile range) | 3.88(.10) | 3.86(.17) | 3.95(.18) | 4.04(.24) |
| actual least squares | 3.88 | 3.90 | 3.89 | 4.11 |

Comparing the students' average slope to the actual slope, we see that the slope of the least squares regression of $y$ on $x$ is close to the average in each data set except F. One possible explanation might be that students tended to fit the

slope of the first principal component or major axis (the line that minimizes the sum of squares of perpendicular rather than vertical distances). The principal component slope is closer to the median slope in every case except N , and is notably closer for F .

Because the y-intercept at $\bar{x}$ is the same for the regression and major axis lines, the conclusion here is simply that the students placed their lines near the centroid of the cloud of points in each case.

By computing the correlation matrix for the students' slopes for the four data sets, we see in Table 2 that students who gave steep slopes for one data set also tended to give steep slopes on the others. This effect seems slight but is definite. The negative values arise because data set N has negative slope.

TABLE 2.
Correlations Between Slope Estimates

| | F | V | N |
|---|---|---|---|
| S | .18 | .14 | -.14 |
| F | | .28 | -.08 |
| V | | | -.05 |

The individual-to-individual variability in slope and in intercept was near the standard error provided by least squares for the four data sets. Using comparable measures of variability, that for slopes was 0.9 times and that for intercepts was 0.7 times the least-squares standard error. Admittedly no theory encourages us to believe in such relations, and further empirical work might be instructive.

## ACKNOWLEDGEMENTS

We wish to express our appreciation to the students for their participation. Nina Leech measured the coordinates of the x's provided by the students. J.W. Tukey suggested the etched line on the transparency. The following provided valuable advice: John Emerson, Katherine Godfrey, Colin Goodall, David Hoaglin, Anita Parunak, Nancy Romanowicz, and Michael Stoto.

## REFERENCES

BEYER, W.H., Editor (1971). Basic Statistical Tables, Cleveland: The Chemical Rubber Co.

DUNN, O.J. and CLARK, V.A. (1974). Applied Statistics: Analysis of Variance and Regression, New York: John Wiley & Sons.

FINNEY, D.J. (1951). "Subjective Judgment in Statistical Analysis: An Experimental Study", Journal of the Royal Statistical Society, Series B, 13, pp. 284-297.